Statistics Coursework

Submit your answers on Moodle by the deadline of 2pm Monday 27th November 2023.

Answers should be a pdf format document (e.g., generated from Word). The pdf document should include any **numerical results** and **plots of results** and **textual responses** to questions and include **ALL the R-code** necessary to achieve the correct answer for each part, i.e., such that a marker could just cut-and-paste and run the code and it would work. The R code sections should be in a uniform width font such as Courier. Always explain your reasoning regarding statistical significance when reaching conclusions.

Always use an individual test significance level of 5% (unless otherwise instructed).

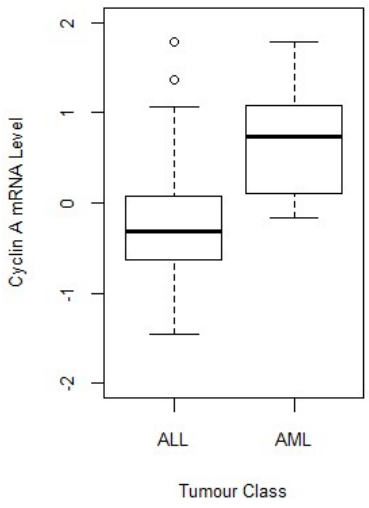
Part A

The five questions in Part A are each worth 10 Marks

Choose and answer THREE questions

1.

a) Write R code to create the boxplot below of the expression levels of the "Cyclin A1 m RNA" from the `golub` dataset. Hint: you can view the gene names with `View(golub.gnames)`. Make the plot a similar as possible to that below.



b) Test the hypothesis that the median expression levels of this gene differ between ALL and AML patients at a significance level $\alpha=0.05$.

2.

It is suspected that mutations in gene X are involved in the survival of cancer patients undergoing a drug treatment. Of 236 patients diagnosed with a particular form of cancer, it is found that 82 have a mutation in gene X and the remainder have the normal version of the gene. All of the patients take the drug for one year, of these 87 die within one year and the rest survive. Of the survivors, 42 have a mutation in gene X.

(a) Organize the data into a contingency table. Include the table in your answer.

(b) Use `chisq.test()` to determine whether survival is or is not dependent on the mutation of gene X. What is your conclusion?

3.

When protein-small molecule complexes form both the protein and the small molecule reduce their surface area that is exposed to water

It is desired to seek evidence of a relationship between this 'buried' surface area and observed binding affinity for 15 small molecule inhibitors binding to DNA gyrase. The surface area buried is split into two parts – non-polar and polar.

```
nonpolar_area <- c(1021, 401, 582, 650, 201, 430, 689, 660, 624,
452, 458, 240, 304, 476, 566)

polar_area <- c(278, 266, 299, 143, 199, 84, 50, 201, 47, 213, 123,
143, 128, 238, 134)

affinity <- c( -53.8, -35.2, -41.9, -47.9, -20.3, -34.2, -46.5, -
43.2, -41.4, -37.6, -40.3, -14.2, -20.1, -36.8, -45.9)
```

(a) Use correlation-based tests to examine the possible relationship between nonpolar area and affinity and between polar area and affinity. What can you conclude from these data?

(b) What were your reasons for selecting the particular statistical test or tests?

4.

The following data describe the levels of an enzyme and a metabolite in a set of combined proteome / metabolome experiments on 20 cultures of *E. coli*

```
enzyme <- c(0.114, 0.510, 0.722, 1.276, 1.928, 2.150, 2.238, 2.732,
2.758, 3.015, 3.616, 3.951, 4.281, 5.315, 6.693, 6.964, 7.056, 8.162
, 8.216, 8.410)

metabolite <- c(11.2, 12.5, 9.5, 12.8, 12.7, 14.8, 18.2, 17.2, 23.4,
24.4, 27.9, 31.8, 34.8, 44.3, 67.1, 72.7, 76.7, 95.1, 100.1, 102.1)
```

a) Find the best-fit polynomial equation that describes the dependence of metabolite level on enzyme level. [Hint: you will need to test at least four models to identify the best fit]

b) Create two plots that demonstrate that your model is a good fit to the data.

5.

We wish to examine the power of the Kolgomorov-Smirnov (KS) test in a particular situation.

a) Define the power of a test.

b) What is the Null hypothesis of the KS test ?

c) By repeatedly creating random samples of size = 100 from a `beta(6,2)` distribution calculate the power of the KS test to identify the difference between the distribution of data from this Beta distribution and from a normal distribution with the same sample mean and sample standard deviation for this sample size.

Part B

The three questions in Part B are each worth 20 Marks

Choose and answer ONE question

6.

Mass spectrometry measurements of the proteome that quantitate the amount of each protein present are known to be NOT normally distributed about the true quantity of protein present. A series of experiments is carried out on wild-type and knockout mutant cell lines. A transcription factor has been deleted in the knockout cell line. For each of the cell lines, 20 replicate measurements of the concentration of a protein X are carried out by mass spectrometry.

```
wildtype <- c(560, 968, 3297, 1200, 858, 646, 992, 2507, 2037, 546, 2929,
1171, 1389, 1958, 3149, 1165, 2257, 2120, 65, 1571)

knockout <- c(589, 232, 983, 2597, 827, 1363, 634, 12, 643, 1889, 2840,
1291, 939, 811, 3290, 525, 90, 543, 2400, 3012)
```

The researchers wish to report the results of these experiments and to determine if the measurements support the idea that deletion of the transcription factor changes the median concentration of protein X present in the cell.

(a) Why is it better to report the results of these experiments in terms of the median value of the measurements rather than the mean?

(b) Use a bootstrap approach to calculate a 95% confidence interval for the median protein X concentration of each cell line.

(c) Use a bootstrap approach to test if the medians of the two cell lines differ.

(d) Suggest and perform another appropriate test for difference between the levels of protein X in these two cell lines.

7. The replication of four mutant strains of influenza virus (types A, B, C, D) is measured in chickens. For each mutant strain, 50 chickens are infected by the virus and the relative amount of virus in each chicken is measured after 24 hours by taking blood samples and measuring the total amount of viral RNA (using a reverse trancriptase and a quantitative polymerase chain reaction). The scientist hypothesizes that the mutations affect replication of the virus in the host.

```
strainA <-c(1.628, 7.054, 1.298, 3.419, 1.815, 5.908, 0.254, 5.475, 1.528,
4.487, 0.370, 0.448, 3.365, 0.401, 1.631, 3.525, 0.440, 5.894, 2.592,
0.584, 1.955, 1.394, 3.595, 0.770, 3.235, 0.031, 2.218, 6.151, 1.026,
5.006, 0.569, 1.451, 0.346, 4.271, 0.150, 2.130, 4.300, 6.327, 8.777,
4.039, 3.006, 1.873, 1.996, 1.173, 3.065, 1.502, 5.678, 5.518, 3.235,
4.185)

strainB <-c(4.106, 3.497, 4.922, 4.419, 1.544, 0.664, 1.736, 6.163, 6.343,
2.581, 4.363, 0.804, 4.915, 4.189, 2.947, 1.661, 2.602, 0.856, 6.236,
3.442, 1.134, 2.711, 1.458, 3.878, 3.294, 2.248, 4.815, 1.908, 0.646,
3.695, 0.263, 0.337, 0.433, 2.424, 1.170, 1.052, 3.906, 1.452, 1.311,
0.113, 3.605, 1.476, 2.376, 6.057, 0.756, 6.298, 7.256, 2.495, 4.224,
0.256)

strainC <-c(3.539, 5.455, 4.782, 2.336, 4.901, 0.895, 5.649, 2.798, 1.110,
10.537, 0.891, 8.877, 3.617, 1.065, 6.362, 2.437, 4.702, 2.982, 3.322,
4.256, 1.854, 4.379, 3.736, 1.745, 3.728, 4.580, 5.211, 3.026, 2.429,
3.922, 2.950, 1.208, 4.520, 2.088, 4.859, 6.910, 2.265, 9.034, 3.512,
6.974, 6.253, 5.897, 4.412, 3.891, 5.562, 1.935, 1.379, 2.822,
5.426, 4.710)

strainD <-c(2.010, 1.734, 0.544, 6.426, 0.171, 6.683, 4.301, 0.194, 8.818,
6.352, 2.338, 1.003, 1.865, 9.068, 2.758, 7.450, 0.801, 2.112, 7.625,
1.095, 4.837, 7.068, 5.707, 0.244, 2.778, 8.518, 0.356, 0.467, 3.794,
0.162, 0.605, 0.725, 2.988, 0.034, 2.022, 1.199, 3.309, 1.696, 0.527,
0.787, 6.263, 4.721, 6.691, 2.334, 2.331, 3.330, 8.584, 7.672, 2.041,
1.168)
```

(a) Test the data for difference in mutation rate. Explain and provide evidence for all your decisions in selecting the test.

(b) You should be able to show in part a) that there is a significant variation among the mutant viruses at the $\alpha = 0.05$ level. Now identify those pairs of strains which have significant differences in replication rate.

(c) If this experiment and analysis in part a) were to be run again many times, in what proportion of repeats would you have reached the conclusion that there is significant variation among the mutant viruses at the $\alpha = 0.05$ level? [Hint: this is problem suitable for a resampling approach]

(d) If the sample size was 20 (instead of 50) in what proportion of repeats would you reach the conclusion that there is significant variation among the mutant viruses at the $\alpha = 0.05$ level?

8. The following data describes the concentrations of three cellular enzymes and a metabolite B in a set 20 experiments.

```
enzymeA <- c(0.114, 0.510, 0.723, 1.276, 1.93, 2.151,2.239, 2.732, 2.759,
             3.015, 3.617, 3.951, 6.281, 6.315, 6.693, 6.965, 7.057, 8.162,
             8.216, 8.410)

enzymeB <- c(0.240, 4.011, 11.915, 12.393, 14.019, 15.047,19.418, 22.978,
            23.483, 29.311, 29.475, 31.647, 32.529, 32.827, 40.212, 44.497,
            44.525, 47.422, 49.717, 49.824)


enzymeC <- c(1.872, 3.177, 3.802, 4.028, 5.050, 5.227, 5.962, 8.211, 9.038,
             9.392, 9.738, 10.981, 12.145, 15.381, 16.897, 17.639, 17.744,
             18.373, 18.692, 19.350)


metaboliteB <- c(28.6, 35.8, 65.7, 135.2, 183.7, 229.2, 221.7, 256.9,
                 274.8, 312.0, 415.0, 490.4, 1014.2, 1085.9, 1222.2, 1308.6,
                 1351.1, 1680.8, 1722.0, 1793.5)
```

a) Investigate the relationships between individual enzyme concentrations and the metabolite level. On the basis of these relationships alone, for which of the enzymes is there evidence to support the idea that they are involved in controlling the level of the metabolite?

b) Find the best-fit multiple-linear regression model for the dependence of metabolite on the enzyme concentrations and justify all your decisions with statistical evidence. [Hint: You will have to create and test many possible models to identify and evidence the best fit.]

c) Create plots that demonstrate that your model is a good fit to the data